Curbing domestic violence: Instantiating C-K theory with Formal Concept Analysis and Emergent Self Organizing Maps

Jonas Poelmans[1], Paul Elzinga[3], Stijn Viaene[1,2], Guido Dedene[1,4]

[1]K.U.Leuven, Faculty of Business and Economics, Naamsestraat 69,
3000 Leuven, Belgium
[2]Vlerick Leuven Gent Management School, Vlamingenstraat 83,
3000 Leuven, Belgium
[3]Amsterdam-Amstelland Police, James Wattstraat 84,
1000 CG Amsterdam, The Netherlands
[4]Universiteit van Amsterdam Business School, Roetersstraat 11
1018 WB Amsterdam, The Netherlands
{Jonas.Poelmans, Stijn.Viaene, Guido.Dedene}@econ.kuleuven.be
Paul.Elzinga@amsterdam.politie.nl

In this paper we propose a human-centered process for knowledge discovery from unstructured text that makes use of Formal Concept Analysis and Emergent Self Organizing Maps. The knowledge discovery process is conceptualized and interpreted as successive iterations through the Concept-Knowledge (C-K) theory design square. To illustrate its effectiveness, we report on a real-life case study of using the process at the Amsterdam-Amstelland police in the Netherlands aimed at distilling concepts to identify domestic violence from the unstructured text in actual police reports. The case study allows us to show how the process was not only able to uncover the nature of a phenomenon such as domestic violence, but also enabled analysts to identify many types of anomalies in the practice of policing. We will illustrate how the insights obtained from this exercise resulted in major improvements in the management of domestic violence cases.

Key words: Formal Concept Analysis, Emergent Self Organizing Map, C-K theory, text mining, actionable knowledge discovery, domestic violence.

---

1. Introduction

In this paper we propose a human-centered process for knowledge discovery from unstructured text that makes use of Formal concept Analysis (FCA) [22,23] and Emergent Self Organizing

Maps (ESOM) [9,12]. Human-centered Knowledge Discovery in Databases (KDD) refers to the constitutive nature of human interpretation for the discovery of knowledge, and stresses the complex, interactive process of KDD as being led by human thought [30]. Data mining should be primarily concerned with making it easy, practical and convenient to explore very large databases for organizations and users with vast amounts of data but without years of training as data analysts [33]. A significant part of the art of data mining is the user's intuition with respect to the tools [15, 16].

Visual data exploration [39] and visual analytics [40] are especially useful when little is known about the data and exploration goals are vague. Since the user is directly involved in the exploration process, shifting and adjusting the exploration goals is automatically done if necessary. In addition to the direct involvement of the user, the main advantages of visual data exploration over automatic data mining techniques from statistics or machine learning are: visual data exploration usually allows a faster data exploration and often provides better results, especially in cases where automatic algorithms fail. In addition, visual data exploration techniques may provide a higher degree of confidence in the findings of the exploration [38].

This paper extends but also synthesizes our previous work involving FCA and ESOM, two visually appealing data exploration aids, for knowledge discovery from unstructured text. In [51], we first discussed the possibilities of using FCA for knowledge discovery in a police environment. A parallel research track consisted of investigating the potential of using ESOM for knowledge discovery. Our first findings using the ESOM are discussed in [52, 17]. We also compared ESOM's performance to that of other SOMs such as the spherical SOM and we found it to be superior [53]. In [41], we briefly presented our idea to use FCA and ESOM together for domestic violence discovery. The ESOM functions as a catalyst for the FCA based discovery process. The

proposed methodology recognizes the important role of the domain expert in mining real-world enterprise applications and makes efficient use of specific domain knowledge, including human intelligence and domain-specific constraints.

We used a semi-automated approach since the major drawback of all automated and supervised machine learning techniques, including decision trees, is that these algorithms assume that the underlying concepts of the data are clearly defined, which is often not the case. These techniques allow almost no interaction between the human actor and the tool and fail at incorporating valuable expert knowledge into the discovery process [38], which is needed to go beyond uncovering the fool's gold [32]. In the paper presented by Hollywood et al. [37] these problems were clearly addressed in the context of terrorist threat assessment. The central question was whether it is possible to find terrorists with traditional automated data mining techniques and the answer was no.

The knowledge discovery process is conceptualized and interpreted as successive iterations through the C-K theory design square. C-K theory offers a formal framework that interprets existing design theories as special cases of a unified model of reasoning [1, 2]. It provides a clear and precise definition of design that is independent of any domain of professional tradition [3]. C-K theory defines design reasoning dynamics as a joint expansion of the Concept (C) and Knowledge (K) spaces through a series of continuous transformations within and between the two spaces. The beauty of C-K theory is that it can provide insight into an iterative and expansive knowledge acquisition process [4, 5]. One of the core characteristics of C-K theory is this focus on human intelligence as the driving force in expanding the space of knowledge. To our knowledge, this is the first systematic application of C-K theory to the information systems domain. C-K theory is used as a unifying framework to provide a clear structure to the discovery

process based on FCA and ESOM. The combined use of FCA and ESOM in the C-K framework not only gives insight into the generic nature of the KDD activity but also makes for significantly improved results. Some of the aspects of this paper have already been discussed in the literature in a fragmented way (e.g. information retrieval, knowledge browsing, prior knowledge incorporation), but an integrated approach has never been pursued.

To illustrate its effectiveness, we report on a real-life case study on using the process at the Amsterdam-Austell and police in the Netherlands aimed at distilling concepts for domestic violence from the unstructured text in filed reports. The aim of our research was to conceptualize and improve the definition and understanding of domestic violence with the ultimate goal of improving the detection and handling of domestic violence cases. One important spin-off of this exercise that will be elaborated on in this paper was the development of a highly accurate and comprehensible classification procedure for automatically raising a domestic violence flag for incoming police reports. This procedure automatically classifies 91% of incoming cases correctly whereas in the past all cases had to be dealt with manually. We performed this classification exercise to measure the quality of our conceptualization of domestic violence. We have never seen a similar set up in the literature and to the best of our knowledge there is no packaged automated solution to do all the same at once.

Over 90% of the information available to police organizations is stored as plain text. To date, however, analyses have primarily focused on the structured portion of the available data. Only recently the first steps have been taken to apply text mining in criminal analysis [6, 20]. Domestic violence is one of the top priorities of the Amsterdam-Amstelland police force in the Netherlands [35]. In the past, intensive audits of the police databases of filed reports established that many of the reports tended to be wrongly labeled as domestic or as non-domestic violence

cases. Previous attempts have mainly focused on developing a machine learning classifier that automatically classified incoming cases as domestic or as non-domestic violence. Unfortunately they were unsuccessful because the underlying concept of domestic violence was never challenged. These systems did not provide any insight into the problem, since they are black-boxes and their classification performance was around 80% only [31]. As a consequence, these systems never made it into operational policing practice. All of these previous attempts had in common that the concept of domestic violence was never challenged. The developers overlooked the complexity of the notion of domestic violence, were unaware that different people have different visions about the nature and scope of it and did not pay attention to niche cases. Moreover, the correctness of the labels assigned to cases by police officers was never verified. We found that different police officers regularly assigned different labels to the same situation. Finally, the developers did not dispose of a high-quality domain-specific thesaurus that contained sufficient discriminant terms for accurately classifying cases. In the past, several automated term extraction and thesaurus building techniques were used [46]. We interviewed several domain experts that were exposed to these efforts. All of them attested to their failure in constructing a useful thesaurus when we asked them for their appraisal of these prior initiatives.

The remainder of this paper is composed as follows. In section 2, we elaborate on the essentials of FCA, ESOM and C-K theory. In section 3, we show how we used the synergistic combination of FCA and ESOM to institute the C-K framework. Section 4 then discusses the dataset, while section 5 showcases the knowledge discovery process and the four C-K operators described in section 2. In section 6, we summarize the actionable results of the iterative knowledge enrichment. Finally, section 7 presents the main conclusions of this paper.

2. FCA, ESOM and C-K theory

2.1 Formal Concept Analysis

FCA arose twenty-five years ago as a mathematical theory [22, 28] and has over the years grown into a powerful tool for data analysis, data visualization and information retrieval [24]. The usage of FCA for browsing text collections has been suggested before by Cole et al. [34]. However, none of the papers have focused on how FCA can be used in an actionable environment for knowledge enrichment and for discovering different types of knowledge in unstructured text. FCA has been applied in a wide range of domains, including medicine, psychology, social sciences, linguistics, information sciences, machine and civil engineering, etc [47]. For instance, FCA has been applied to analyzing data of children with diabetes [48], for developing qualitative theories in music esthetics [49], for database marketing [49], and for an IT security management system [50]. In [42, 43], FCA was used as a visualization technique that allows human actors to quickly gain insight by browsing through information.

We previously applied FCA to a relatively small police dataset and were able to establish its practical usefulness [51]. FCA is particularly suited for exploratory data analysis because of its human-centeredness [44, 45]. It is a fundamental principle that the generation of knowledge from information is promoted by representations that make the inherent logical structure of the information transparent. FCA builds on the model that concepts are the fundamental units of human thought. Hence, the basic structures of logic and logical structure of information are based on concepts and concept systems [26, 27]. Consequently, FCA uses the mathematical abstraction of the concept lattice to describe systems of concepts to support human actors in their information discovery and knowledge creation practice [25]. The details of FCA theory can be found in Appendix A.

2.2 Emergent Self Organizing Map

Emergent Self Organizing Maps (ESOM) [9] are a special and very recent type of topographic maps [7, 8, 14]. According to [12], "emergence is the ability of a system to produce a phenomenon on a new, higher level". In order to achieve emergence, the existence and cooperation of a large number of elementary processes is necessary. An Emergent SOM differs from a traditional SOM in that a very large number of neurons (at least a few thousands) are used [10]. In the traditional SOM, the number of nodes is too small to show emergence. ESOM is argued to be especially useful for visualizing sparse, high-dimensional datasets, yielding an intuitive overview of their structure [13]. From a practitioner's point of view, topographic maps are a particularly appealing technique for knowledge discovery in databases [13, 19] because they perform a non-linear mapping of the high-dimensional data space to a low-dimensional space, usually a two-dimensional one, which facilitates the visualization and exploration of the data [11]. In the past, we applied the ESOM to a police dataset and found its performance to be superior to that of a spherical SOM tool [54]. We made some interesting discoveries using the ESOM, although the obtained results were limited and not convincing enough to make it into operational policing practice [52].

It is claimed by Ultsch and co-workers that the topology preservation of the traditional SOM projection is of little use when the maps are small: the performance of a small SOM is argued to be almost identical to that of a *k*-means clustering, with *k* equal to the number of nodes in the map [9]. Using large numbers of neurons, as in the ESOM, permits one to observe data at a higher level capturing the overall structures, disregarding the elementary ones and allowing the consideration of structures that otherwise would be invisible. The details of ESOM theory can be found in Appendix B.

2.3 C-K theory

The Concept-Knowledge theory (C-K theory) was initially proposed and further developed by Hatchuel et al. [1, 2, 3, 4, 5]. C-K theory is a unified design theory that defines design reasoning dynamics as a joint expansion of the Concept (C) and Knowledge (K) spaces through a series of continuous transformations within and between the two spaces [4]. C-K theory makes a formal distinction between Concepts and Knowledge: the knowledge space consists of propositions with logical status (i.e. either true or false) for a

designer, whereas the concept space consists of propositions without logical status in the knowledge space. According to Hatchuel et al. [4], concepts have the potential to be transformed into propositions of K but are not themselves elements of K. The transformations within and between the concept and knowledge spaces are realized by the application of four operators: concept $\rightarrow$knowledge, knowledge $\rightarrow$concept, concept $\rightarrow$concept and knowledge $\rightarrow$ knowledge. These transformations form what Hatchuel calls the design square, which represents the fundamental structure of the design process. The last two operators remain within the concept and knowledge spaces. The first two operators cross the boundary between the Concept and Knowledge domains and reflect a change in the logical status of the propositions under consideration by the designer (from no logical status to true or false, and vice versa).

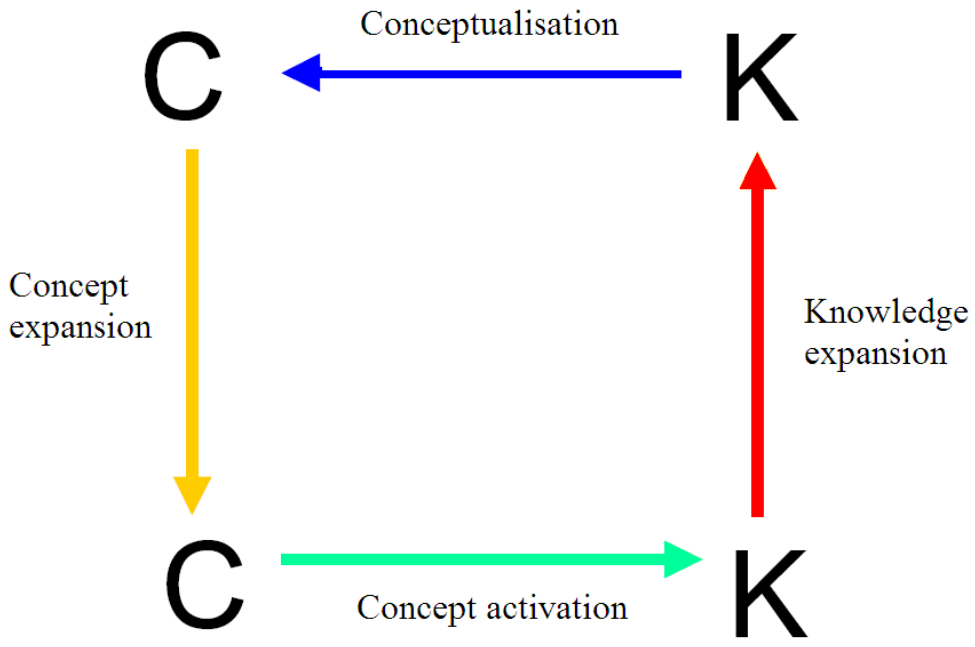

Fig. 1. Design square (adapted from [4])

Design reasoning is modeled as the co-evolution of C and K. Proceeding from K to C, new concepts are formed with existing knowledge. A concept can be expanded by adding, removing or varying some attributes (a "partition" of the concept). Conversely, moving from C to K, designers create new knowledge either to validate a concept or to test a hypothesis, for instance through experimentation or by combining expertise. The iterative interaction between the two spaces is illustrated in Figure 1. The beauty of

C-K theory is that it offers a better understanding of an expansive process. The combination of existing knowledge creates new concepts (i.e. conceptualisation), but the activation and validation of these concepts may also generate new knowledge from which once again new concepts can arise.

However, one of the reasons why it is hard to apply traditional C-K theory in practice is that it lacks an actionable definition of the notions concept, partition and inclusion. In this paper, we show that these issues can be resolved by implementing the C-K framework with a synergistic combination of FCA and ESOM for modeling and expanding the space of concepts. One of the limitations of traditional C-K theory is that hierarchical representations are used to model and expand the concept space. These hierarchical representations are limited in their semantic expressiveness, which is one of the reasons why we chose for the non-hierarchical concept representation of FCA. Complementary to FCA, the ESOM functions as a catalyst to make the knowledge discovery process with FCA more efficient. One of the issues we encountered while using FCA was the scalability of the techniques for larger datasets. We choose to solve this problem by using the ESOM maps, which provide a clear picture of the overall distribution of the entire dataset and the available clusters. The combination of the maps and lattices allows for an efficient exploration of the data, leading, amongst other things, to a better selection of police reports for in-depth manual inspection.

3. Instantiating C-K theory with FCA and ESOM

In this section, we elaborate on the applied process for knowledge discovery based on the visually appealing discovery techniques presented in section 3. FCA as a standalone technique suffers from scalability issues when the number of attributes is increased. Exploring high-dimensional data and discovering new concepts with FCA while little is known about the contents is a difficult task. Although the ESOM can provide some insights into the overall distribution of the data and may help in discovering new concepts and knowledge in the data, its capacities for knowledge discovery are limited. The ESOM as a standalone technique does not allow gaining thorough insights into the conceptual structure of the data and the underlying knowledge of police officers. This is important since we want to improve our understanding of the

gaps in the current domestic violence definition, the knowledge of police officers concerning the problem, etc. In this paper, we go beyond the use of either one of these techniques and use them in combination as part of a unifying framework based on C-K theory. The unifying framework gives insight into the generic nature of the KDD activity and is a necessary precondition for successfully embedding the knowledge discovery process based on the synergistic combination of FCA and ESOM in daily policing practice. In this setup, FCA is used as a concept engine, distilling formal concepts from unstructured text. We complement knowledge discovery with the capabilities of ESOM, which functions as a catalyst for the FCA based knowledge extraction. Our approach to knowledge discovery is framed in C-K theory. The K space could be viewed as being composed of actionable information. It contains the existing knowledge used to operate and steer the action environment. The C space, on the other hand, can be considered as the design space. Whereas K is used as the basis for action and decision making, C puts this actionability under scrutiny for potential improvement and learning. At the basis of the knowledge discovery process are many fast iterations through the C-K loop.

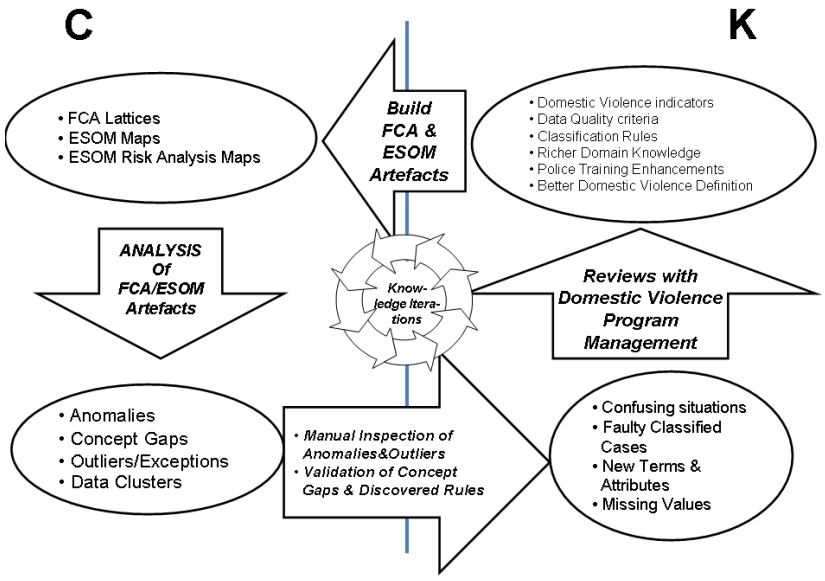

Fig. 4. Knowledge discovery process

During the mining process, two persons, an exploratory data analyst and a domain expert are the driving force behind the exploration and collaborate intensively. There is a continuous process of iterating back and forth between the FCA lattices, the ESOM maps and the police reports. The knowledge discovery process using the combination of FCA and ESOM is summarized in Figure 4. It basically consists of iteratively applying the four operators from the design square in Figure 1.

Initially, an FCA lattice and an ESOM map are constructed by the exploratory data analyst based on the domain expert's prior knowledge of the problem area, the police reports contained in the dataset and the terms contained in the thesaurus (i.e. K $\rightarrow$ C). The lattice and the ESOM map provide a reduced search space to the domain expert, who then visually inspects and analyzes the lattice and ESOM map (i.e. C $\rightarrow$ C). The synergistic combination of FCA and ESOM can be considered as a knowledge browser. Our contention is that it allows for an effective interaction between the human actors and the underlying information. Using FCA, police reports are selected for in-depth manual inspection based on observed anomalies and counter-intuitive facts (i.e. C $\rightarrow$ K). Using the ESOM map, police reports are selected based on the analysis of outliers, clusters and areas of the map containing a mixture of domestic and non-domestic violence cases (i.e. C $\rightarrow$ K). These police reports are then used to discover new referential terms to improve the thesaurus, to enrich and validate prior domain knowledge, to discover new classification rules or for operational validation (i.e. K $\rightarrow$ K).

Additionally, based on the classification rules discovered using FCA, we label/relabel cases and use these cases to construct an ESOM risk analysis map. We then project the unlabeled cases onto this map (i.e. K $\rightarrow$C). Subsequently, this map is analyzed by the exploratory data analyst and the domain expert, who search the map for outliers, clusters of cases in different areas of the map and areas containing a mixture of domestic and non-domestic violence cases (i.e. C $\rightarrow$ C). Based on the observations made, representative police reports are again selected for in-depth manual inspection (i.e. C $\rightarrow$ K). The obtained results, together with the relevant prior knowledge of the domain expert, are then incorporated into the existing visual representation, resulting in a new lattice and ESOM map (i.e. K $\rightarrow$C).

4. Dataset

Our dataset consists of a selection of 4814 police reports describing a whole range of violent incidents from the year 2007. All domestic violence cases from that period are a subset of this dataset. The selection came about amongst others by filtering out those police reports that did not contain the reporting of a crime by a victim, which is necessary for establishing domestic violence. This happens, for example, when police officers are sent to investigate an incident and afterwards write a report in which they mention their findings, but the victim ends up never making an official statement to the police. The follow-up reports referring to previous cases were also removed. From the 4814 police reports contained in the dataset the following information was extracted: the person who reported the crime, the suspect, the persons involved in the crime, the witnesses, the project code and the statement made by the victim to the police. Of those 4814 reports, 1657 were classified by police officers as domestic violence. These data were used to generate the 4814 html-documents that were used during our research. An example of such a report is displayed in Figure 5.

The validation set for our experiment consists of a selection of 4738 cases describing a whole range of violent incidents from the year 2006 where the victim made a statement to the police. Again, the follow-up reports were removed. Of these 4738 cases 1734 were classified as domestic violence by police officers.

Title of incident	Violent incident xxx
Reporting date	31-03-2008
Project code	Domestic violence against ex-partner
Crime location	Amsterdam Wibautstraat yyy
Suspect (male) Suspect (18-45yr)	Zzz
Address	Amsterdam Waterlooplein yyy
Involved (male) Involved (>45yr)	Neighbours
Address	Amsterdam Wibautstraat www
Victim (female) Victim (18-45jr)	Uuu
Address	Amsterdam Waterlooplein vvv

Reporting of the crime

Yesterday morning I was taking a bath. Suddenly my daughter ran into the bathroom followed by her ex-boyfriend. She screamed for help. He had a gun in his hand and he was clearly under the influence of beer or drugs. He yelled out that he couldn't live without her. He threatened to kill me and my daughter if she wouldn't come back to their house. The neighbors who were alarmed by all the noise came to lend some help. Meanwhile another neighbor phoned the police. I jumped out of the bath and tried to push him on the floor. During this fight I got some serious injuries on my back etc.

Figure 5. Example police report

The initial phase of the knowledge acquisition process consists of translating the area under investigation into objects, terms and attributes. We considered the police reports from the dataset as objects and the relevant terms contained in these reports as attributes. The terms and term clusters (see section 6) are stored in a thesaurus.

We composed an initial thesaurus of which the content was based on expert prior knowledge such as the domestic violence definition. We enriched the thesaurus with terms referring to the different components of the definition such as "hit", "stab", "my mother", "my ex-boyfriend". Since domestic violence is a phenomenon that according to the literature typically occurs inside the house, we also added terms such as "bathroom", "living room". We made an explicit distinction from public locations such as "under the bridge", "on the street". The initial thesaurus contained 123 elements.

The reports were indexed using this thesaurus. For each report the thesaurus elements that were encountered were stored in a collection. This collection would be used as input for both the FCA and the ESOM procedure. The thesaurus was refined after each iteration of re-indexing the reports and visualizing and analyzing the data with the FCA lattice and ESOM maps. This process is demonstrated in detail in section 6.

5. Iterative knowledge discovery with FCA and ESOM

In this section, we illustrate the abstract description of the knowledge discovery process provided in section 3 with a real life case study with the Amsterdam-Amstelland police on domestic violence. We have chosen not to present the sequential build-up of the lattices and ESOM maps, but to make a selection

from these lattices and maps, just to help the reader become familiar with the explorative possibilities of the method presented here.

The process displayed in Figure 4 contains an iterative learning loop. During the successive iterations through the C-K loop, multiple interesting results emerged from the research. These different types of results will now briefly be described. The analysis process is showcased in detail in the next subsections. The FCA lattices and ESOM maps are mainly used as an instrument to efficiently select representative reports for in-depth manual inspection, to discover new classification rules, to enrich, test and refine expert prior knowledge, to browse and annotate the collection of police reports, etc.

An important aspect of the process consists in searching these reports for new attributes that can be used to discriminate between the domestic and non-domestic violence reports or that may lead to an enrichment of existing domain knowledge. New referential terms were not selected using a term extractor, but they were obtained by carefully reading some representative reports and then selecting relevant terms as attributes. We built in the necessary validation mechanisms to ensure the completeness of the thesaurus:

1. Word stemming. Each word is reduced to word-stem form.
2. Stop wording. A stop list is used to delete from the texts the words that are insufficiently specific to represent content. The stop list contains many common function words, such as “the”, “or”, etc.
3. Synonym lists. Synonym lists are used to add semantically similar words.
4. Spelling checking. Spelling checking is used to validate the correctness of the term added to the thesaurus and the correctness of the words in the police reports.

During the research the thesaurus was under constant evolution: when new terms and concepts were discovered, the terms were added to the thesaurus. This approach ensured that the thesaurus remained at all times a reflection of the knowledge already gained. Because of the large number of police reports in the dataset, it was not possible to visually analyze concept lattices containing more than 14 attributes. There-

fore, terms with a similar semantic meaning or referring to the same domain concept were clustered by the domain experts. When these term clusters were used to create an FCA lattice, they were considered as attributes.

During the exploration, we also verified the correctness of the labels assigned by police officers to the selected cases and we searched the reports for new interesting concepts, inconsistencies, etc. This led amongst others to the discovery of faulty case labelings and situations that were often not recognized by police officers as domestic or as non-domestic violence. This information was used by the data quality management team to significantly improve the quality of the data in the police databases and to improve the way police officers handle domestic violence cases. The information was also useful for the domestic violence program manager to improve the training of police officers. We also found some regularly occurring confusing situations that could not be uniquely classified as domestic or non-domestic violence based on the domestic violence definition. These situations were presented to the program manager and were used to enrich, improve and refine the concept and definition of domestic violence.

During the discovery and conceptualization of the nature of domestic violence from the data at hand, we were able to define a set of accurate and comprehensible classification rules to automatically classify incoming cases as domestic or as non-domestic violence. In the past developing an accurate classifier using decision trees, SVMs, Neural Networks, etc. turned out to be impossible. We found that this was largely due to the incorrect labels assigned by police officers to cases, to the vagueness of the domestic violence definition and to the lack of a high-quality thesaurus. We managed to resolve many of these problems during the exploration with FCA and ESOM, resulting in a set of highly accurate and comprehensible classification rules. All these different aspects of the process, which have only been briefly introduced so far, are discussed more extensively in the next sections.

5.1 Transforming existing knowledge into concepts

The process of design reasoning starts by making the transition from the knowledge space to the concept space. The process of transforming propositions of K into concepts of C is called disjunction. The corresponding operator in the design square from Figure 1 is the knowledge $\rightarrow$ concept operator. This operator expands the space of C with elements from K. We used two techniques to perform this knowledge to concept transformation. First, we constructed an FCA lattice based on expert prior knowledge, the police reports in the dataset and the term clusters in the thesaurus. Second, we designed an ESOM map based on the terms in the thesaurus and the police reports in the dataset. Both methods are further discussed in this section.

The definition of domestic violence employed by the police organization of the Netherlands is as follows:

> "Domestic violence can be characterized as serious acts of violence committed by someone in the domestic sphere of the victim. Violence includes all forms of physical assault. The domestic sphere includes all partners, ex-partners, family members, relatives and family friends of the victim. The notion of family friend includes persons that have a friendly relationship with the victim and (regularly) meet with the victim in his/her home [21, 29]".

The lattice in Figure 6 was fundamentally influenced by this domestic violence definition. Prior to the analysis with FCA, certain terms were clustered in term clusters based on this definition and added to the thesaurus. We clustered the terms contained in the thesaurus into term clusters associated with one of the two components of the definition (i.e. prior knowledge incorporation).

We first attempted to verify whether a report could be classified as domestic violence by checking it for the occurrence of one or more terms related to each of the two components of the domestic violence definition. In other words, a case would be labeled as domestic violence if the following two conditions

were fulfilled. First, a criminal offence had occurred. To verify whether a criminal offence had occurred, the report was searched for terms such as “hit”, “stab” and “kick”. These terms were grouped into the term cluster “acts of violence”. Second, a person in the domestic circle of the victim was involved in the crime. Therefore, the report was searched for terms such as “my dad”, “my ex-boyfriend” and “my uncle”. These terms were grouped into the term cluster “persons of domestic sphere”.

Using the reference definition of domestic violence employed by the police was but one way to identify term clusters to structure the lattice in Figure 6. Term clusters also emerged from in-depth scanning of certain reports highlighted during a knowledge iteration cycle. This is how, for example, the term cluster “relational problems” was created. We discovered terms such as “relational problems”, “I had a relationship with”, which refer to a broken relationship. In the analysis of some of the reports selected using ESOM during an earlier iteration, we also found that many cases did not have a formally labeled suspect. This attribute is also incorporated in the lattice in Figure 6.

Reports that were assigned the label “domestic violence” have been classified as such by police officers. The remaining reports were categorized as non-domestic violence. This results in the lattice displayed in Figure 6.

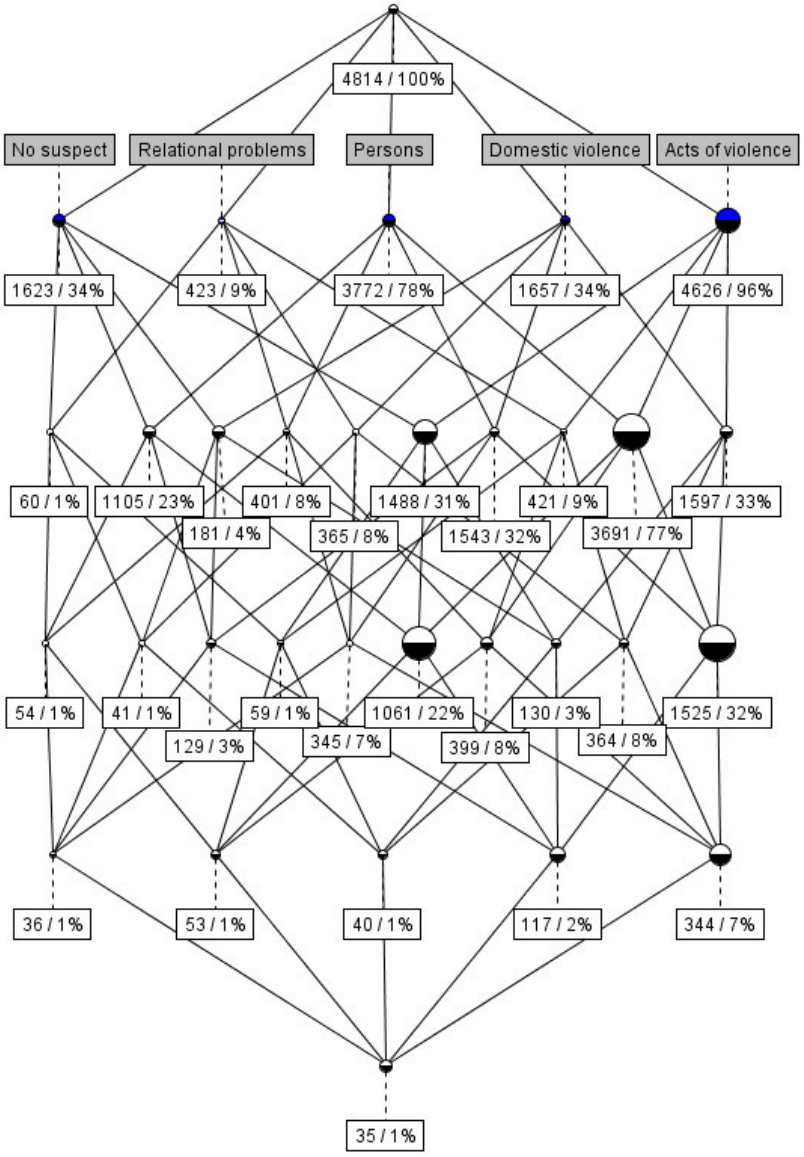

Fig. 6. Initial lattice based on the police reports from 2007

Indexing the 4814 reports from 2007 with the initial thesaurus from section 5 resulted in a cross table with all reports as objects and all terms as attributes. This cross table is used for training a toroidal ESOM. The ESOM is represented in Figure 7: the green squares refer to neurons that dominantly contain non-domestic violence cases, while the red squares refer to neurons that dominantly contain domestic violence cases.

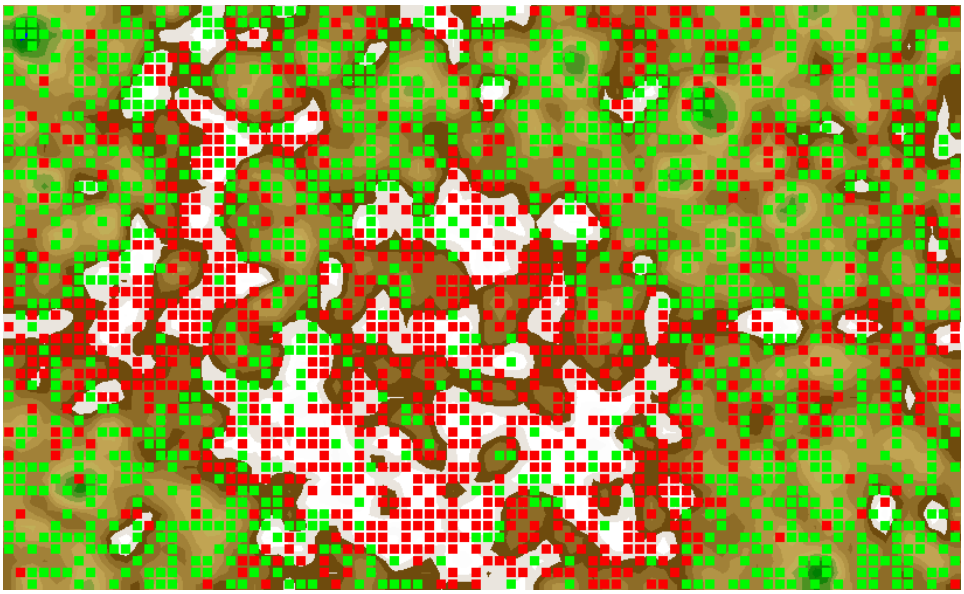

Fig. 7. ESOM map

5.2 Expanding the space of concepts

The notion of expansion plays a key role in C-K theory. An analyst's ability to recognize an expansion can depend on his sensitivity to these opportunities, his training or the knowledge at his disposal. In [5] it is stated that expansion is a K-relative notion, which means that its significance depends on the knowledge of a designer or any other observer or user. In this paper, we argue that FCA and ESOM help analysts recognize and exploit these opportunities. Basically, C space expansion is driven by the analyst's detection and investigation of anomalies, outliers, clusters and concept gaps with these visual exploration tools. Based on these observations, police reports are selected for in-depth manual inspection. This section describes in more detail these two ways of expanding the space of concepts.

We first explain how we used FCA to expand the space of concepts. FCA was used to efficiently explore the data based on the prior knowledge of the domain expert. Some interesting findings emerged from the interactive exploration of the lattice in Figure 6 and warranted further investigation.

Table 2. Interesting observations from the lattice in Figure 6

	Non-domestic violence	Domestic violence

No “acts of violence”	128	60
No “acts of violence” and “persons of domestic sphere”	63	18
“Acts of violence” and no “persons of domestic sphere”	863	72
“Relational problems”	58	365
“No suspect”	1442	181

As can be seen from Table 2, a total of 60 domestic violence cases did not contain a term from the “acts of violence” term cluster. Of these 60 cases 18 contained a term from the clusters containing terms referring to a person in the domestic sphere of the victim. Interestingly, some 28% (i.e. 863) of the non-domestic violence reports only contain terms from the “acts of violence” cluster, while there are only 72 domestic violence reports in the dataset that share that characteristic. Apparently, some cases that were labeled as domestic violence did not fit the definition of domestic violence that was used to start this discovery exercise in the first place. The reports in question were therefore selected for in-depth investigation.

It should be clear from the lattice in Figure 6 that the terms contained in the cluster “relational problems” tend to be associated with domestic violence cases. Apparently, only 58 non-domestic violence reports contained one or more terms from the “relational problems” cluster. We concluded that the presence of at least one of the terms from of this cluster in a police report seemed to be a strong indication for domestic violence. This was enough evidence to warrant manual inspection of these 58 police reports.

Visual inspection of the patterns produced by the ESOM map in Figure 7 also allowed us to make some interesting observations. For example, colour coding made it easy to detect outlying observations: some red squares are located in the middle of a large group of green squares and vice versa. For further examination we made use of the ESOM tool’s functionality to select neurons and display the cases that had this neuron as their best match. We thought that these neurons were associated with cases that might have been wrongly classified by police officers. Therefore, these cases were also selected for in-depth manual inspection.

5.3 Transforming concepts into knowledge

The concept $\rightarrow$ knowledge operator from Figure 1 transforms concepts in C into logical questions in K. In our case an answer to such a question is found by manually inspecting the selected police reports. We refer to this manual analysis as the validation of concept gaps, giving rise to multiple types of discoveries: confusing situations, new referential terms, faulty case labelings, niche cases and data quality problems.

For example, and with reference to Table 1, the 18 cases labeled by police officers as domestic violence that contained a term from the "persons of domestic sphere" but no violence term were selected for manual inspection. Is it possible that there are domestic violence reports in which the victim does mention a person of the domestic sphere, but does not mention an act of violence? In-depth analysis showed that these 18 reports contained violence related terms that were originally lacking from the initial thesaurus, such as "abduction", "strangle" and "deprivation of liberty". Another example is the discovery of 42 cases that did not contain a violence term or a term referring to a person of the domestic sphere. These cases turned out to be wrongly classified as domestic violence. We also analyzed the reports that contained a violence term but no term referring to a person of the domestic sphere. This inspection revealed that more than two thirds of these reports were wrongly classified as domestic violence. In the next section, we will focus on the causes of these labeling errors and the extraction of actionable intelligence from these individual cases that can be used to improve the domestic violence definition and the training of police officers.

Table 1 also indicates that there were 58 police reports that were classified as non-domestic violence while containing a term from the "relational problems" cluster. This investigation revealed that a startling 95% of these cases had been wrongly labeled as non-domestic violence. Moreover, about 70% of these cases had as a common feature that a third person made a statement to the police for someone else. Analysis of the remaining 30% of these misclassified cases led to the discovery of an important new concept that was lacking from the domain expert's initial definition of domestic violence. Many of the reports included expressions such as "I was attacked by the ex-boyfriend of my girlfriend" and "I was harassed

by the ex-girlfriend of my boyfriend". These terms were grouped into the cluster "attack by ex-person against new friend". This situation is analyzed in detail in the next section together with the resulting actionable intelligence. The term cluster is also used to distil new classification rules in one of the subsequent iterations.

Another interesting finding emerged from our search for novel and potentially interesting classification attributes. The lattice in Figure 6 shows that some 34% of the reports (1623 cases) did not mention a suspect. According to the domestic violence definition (which specifies that the perpetrator must belong to the domestic circle of the victim), the offender has to be known in domestic violence cases. Naturally, we had assumed that these reports described non-domestic violence cases. Nevertheless, when looking into these cases, we found that 181 of them turned out to describe domestic violence cases after all. In the next section, we uncover the causes of this phenomenon.

After an in-depth manual inspection of the police reports corresponding to the ESOM outliers, interesting discoveries were made. For example, we observed that many of these outlier reports contained several important new features that were lacking in the domain expert's understanding of the problem area. Every time new and important features were discovered in this way, they were used to enrich the thesaurus. A selection of these features is displayed in Table 3 and 4.

Table 3. Newly discovered features by studying the domestic violence outliers in the ESOM map.

Pepper spray
Homosexual relationship, lesbian relationship
Sexual abuse, incest
Alternative spelling of some words (e.g. ex-boyfriend, exboyfriend, ex boyfriend)
Weapons lacking in the thesaurus: belt, kitchen knife, baseball bat, etc.
Terms referring to persons: partner, fiancée, mistress, concubine, man next door, etc.
Terms referring to relationships: romance, love affair, marriage problems, divorce proceedings, etc.
Reception centers: woman's refuge center, home for battered woman, etc.
Gender of the perpetrator: mostly male
Gender of the victim: mostly female
Age of the perpetrator: mostly older than 18 years and younger than 45 years
Age of the victim: mostly older than 18 years and younger than 45 years
Terms referring to an extra marital affair: I have an another man, lover, I am unfaithful, etc.

Table 4. Newly discovered features by studying the non-domestic violence outliers in the ESOM map.

Places of entertainment: club, disco, bar, etc.
Crime locations: on the street, on a bridge, under a viaduct, on a crossing, etc.
Public locations: metro station, bus stop, tram stop, etc.
Reception centers: refugee center, shelter for the homeless, relief center, etc.
Drugs: drug abuse, drug joint, etc.
Addresses of youth institutions, prisons, etc.
Hotel: hotel room, hotel, etc.
Description of suspect's origin: Turkish descent, white man, North-African descent, etc.
Description of suspect's body: 1.75 meters tall, 119 centimeters tall, muscular appearance, etc.
Description of suspect's hair: curly haired, blond hair, redhead, etc.
Description of suspect's clothes: black jacket, leather shoes, blue pants, jeans, etc.
Description of suspect's face: beard, moustache, facial hair, etc.
Description of suspect's accent
Unknown person is involved in the crime
Attack by unknown person
Corporate body
Neighborhood quarrel

The reports also contained multiple confusing situations. When more detailed information was disclosed to us, these cases were also used to refine the domestic violence definition.

5.4 Expanding the space of knowledge

The expansion of the space K constitutes validation or testing of the proposed expansion with the ultimate goal of producing actionable intelligence. K-validation of a concept boils down to a confrontation of the output from the C-K transformation with knowledge sources available to the K space (e.g. cross-checking with other databases, setting up field experiments, soliciting expert advice). These new propositions have logical status. In this section, we show how we obtain actionable intelligence from the observations made during the Concept $\rightarrow$ Knowledge phase.

Analysis of the misclassified police reports described in the previous section revealed that for some unknown reason police officers regularly seem to misclassify burglary, car theft, bicycle theft and street robbery cases as domestic violence. Therefore, terms such as "street robbery", burglary" and "car theft" were combined into a new term cluster called "burglary cases". This term cluster was then used in one of the subsequent iterations through the C-K loop.

In the previous section, we also described how the analysis of the police reports revealed that a situation in which a third person makes a statement for somebody else can be confusing for police officers. For example, one case described a father who made a statement to the police about the sexual abuse of his daughter by her stepfather. This is a clear case of domestic violence, but since it was not the victim who made the statement to the police, the police officer did not recognize it as such. This type of situation is now specifically addressed in police training.

In the previous section, we also described how the analysis of police reports revealed interesting cases in which the ex-boyfriend attacked the new boyfriend. We presented these ambiguous cases to the board members responsible for the domestic violence policy. Police officers and policy makers confirmed that this type of situation was to be seen as domestic violence, mainly because the perpetrator often intends to emotionally hurt the ex-partner. Consequently, the expectation was for the terms contained in this cluster to frequently occur in domestic violence reports. However, this turned out to be incorrect. It became clear from the investigation that in general this type of situation was very confusing to police officers. A quick scan revealed that more than 50% of police officers actually had trouble with such cases. The ensuing investigation and discussions with police officers and policy makers revealed that this situation needed to be addressed during the training of police officers. Several interesting cases like the previous one were identified during the data exploration. All of them resulted in a clearer insight into the nature of domestic violence.

In the previous section, we found that some domestic violence cases did not mention a formally labeled suspect. Analysis revealed that this was a result of police officers' rather haphazard ways of registering suspects for these cases. Apparently, while some officers immediately registered a suspect at the moment the victim mentioned this person as a suspect, others preferred to first interrogate these supsects before officially labeling them as such. In the latter case, the person would just be added to the list of persons who were said to be involved in or to have witnessed the crime. Because such lists included friends, family members or bystanders, they could potentially be very extensive and diverse, which is why sus-

pects easily got lost in these lists. When we inquired about the proper policy regarding the labeling of suspects, we were told there simply was none. Our analysis made a strong case for the need for such a policy. In the end, the quick-win proposal that could be implemented to solve this issue involved a relatively simple change to the registration software: an additional data entry field would need to be introduced for police officers to register the persons that were mentioned by the victim as offenders.

The newly obtained knowledge led to a new iteration of the FCA analysis, supported by another run of the ESOM tool. In each iteration, it is possible that one or more new classification rules are discovered. The attribute "corporate body", for example, was found by first analyzing a cluster of green squares that was located within a group of red squares in an ESOM map. With FCA we found that the presence of a corporate body in a police report almost always excludes domestic violence. Therefore, we introduced a new domestic violence classification rule named "corporate body".

6. Actionable results

Several iterations through the design square resulted in truly valuable upgrades of the K space from the perspective of improving action in the field. This section provides an overview of some of the most important achievements of our work.

First, we were able to refine the definition of domestic violence that would act as a principle guideline for labeling cases. During the exploration, several types of niche cases were identified as valid exceptions to the general definition. No clear labeling guidelines were available, so we formulated advice, grounded in evidence, to redesign the general policy. Eventually, we obtained the classification guidelines displayed in Table 5.

Table 5. Classification guidelines for incidents involving inhabitants of the same institution

Perpetrator	Victim	Classification
Caretaker	Inhabitant	Domestic violence
Inhabitant	Caretaker	Non-domestic violence
Inhabitant younger than 18y	Inhabitant younger than 18y	Domestic violence
Inhabitant older than 18y	Inhabitant older than 18y	Non-domestic violence
Inhabitant of prison older than 18y	Inhabitant of prison older than 18y	Individual evaluation

Inhabitant older than 18y	Inhabitant younger than 18y	Domestic violence
Inhabitant younger than 18y	Inhabitant older than 18y	Individual evaluation

In the end the presence or absence of a dependency relationship between the perpetrator and the victim was the decisive factor for classifying a case as either domestic or as non-domestic violence. Nevertheless, we also discovered some regularly occurring situations in which there is a clear dependency relationship between the perpetrator and the victim, but that were typically classified as non-domestic violence by police officers. A selection of these circumstances is listed in Table 6. These confusing situations helped to expose the mismatch between the management's conception of domestic violence and that of police officers. We found that the management employed a much broader definition of domestic violence than most police officers.

Table 6. Circumstances in which the offender abuses the dependency relationship with the victim, but that are not recognized by police officers as domestic violence.

Circumstance	Dependency relationship
Lover boys	The victim is in love with the lover boy, who abuses this dependency relationship to force her into prostitution.
Extramarital relationship	If the mistress of an adulterer blackmails him, for example by threatening to reveal their affair to his wife, the mistress abuses the dependency relationship that exists between her and the man.
Violence between a caretaker and an inhabitant of an institution	If the caretaker threatens or harasses the inhabitant (for example, a nurse who maltreats an elderly woman in a retirement home), the latter is often helpless because she depends on the caretaker.
Violence between colleagues	If two colleagues had a relationship and one keeps stalking the other, this is domestic violence between ex-persons.
An ex-boyfriend attacks the new boyfriend	This is considered to be domestic violence because the ex-boyfriend often intends to emotionally hurt his ex-girlfriend.
Third person makes statement to the police for somebody else	Police officers regularly fail to recognize cases in which a third person makes a statement to the police for somebody else (e.g. a father who makes a statement about the sexual abuse of his daughter by her stepfather) as domestic violence.

Second, a set of 22 domestic violence and 15 non-domestic violence classification rules were extracted. Using these rules, 75% of cases from the year 2007 could be labeled automatically as either domestic or non-domestic violence. We also applied these rules to two validation sets containing unstructured police reports from the year 2006 and from the year 2008, which yielded similar results, i.e. 72% and 73% respectively. These rules are now fully operational and used to automatically and correctly clas-

sify the majority of incoming cases, while in the past all cases had to be dealt with manually. Ten of these domestic violence and five of these non-domestic violence classification rules are displayed in Table 7.

Table 7. Excerpt of discovered classification rules

	Domestic violence classification rules
1	Legal proceedings against domestic sphere
2	Committed by domestic sphere
3	Relational problems and living together
4	Relational problems and institutions
5	Honor related violence
6	Incest
7	(Court) injunction
8	Fear of domestic sphere
9	Attack by ex-person against new friend
10	Problems with domestic sphere
	Non-domestic violence classification rules
1	Unknown perpetrator
2	Corporate body
3	Burglary cases
4	Road rage
5	Violence at school

Third, the set of newly identified classification rules did not just allow the police to classify incoming cases. The rules could also be employed to reclassify cases from the past to result in more correct performance management and reporting over time. Domestic violence cases that were not recognized as such in the past might also be re-opened for investigation. In total, we found 420 filed reports that were wrongly labeled as domestic violence and 912 filed reports that were wrongly labeled as non-domestic violence. Table 8 presents an overview of these results.

Table 8. Number of filed reports that were incorrectly classified, but corrected by means of the 37 rules

	Non-domestic corrected to Domestic	Domestic corrected to Non-domestic	Total
Year 2006	307	136	443
Year 2007	290	115	405
Year 2008	315	169	484
Total	912	420	1332

Finally, based on the cases that could be labeled using the classification rules that were discovered, we constructed an ESOM risk analysis map. For each neuron, the number of domestic and non-domestic violence cases contained in the neuron and the 32 surrounding neurons was counted and used to calculate the probability that a police report that has this neuron as its best match described a domestic violence incident. For the visualization, a color scheme consisting of 5 different colors was used. Red indicates a 90-100% probability rate of domestic violence, orange a 70-90% probability rate, yellow a 30-70% probability rate, green a 10-30% probability rate and dark green a 0-10% probability rate. The labels of the cases that could not be categorized using the new classification rules were not used to construct this risk analysis map. However, we projected these remaining cases onto this map afterwards. The map for the dataset of the year 2007 is shown in Figure 8. Cases that were labeled by police officers as domestic violence are represented as black dots, while the cases that were labeled as non-domestic violence, are represented as light blue dots.

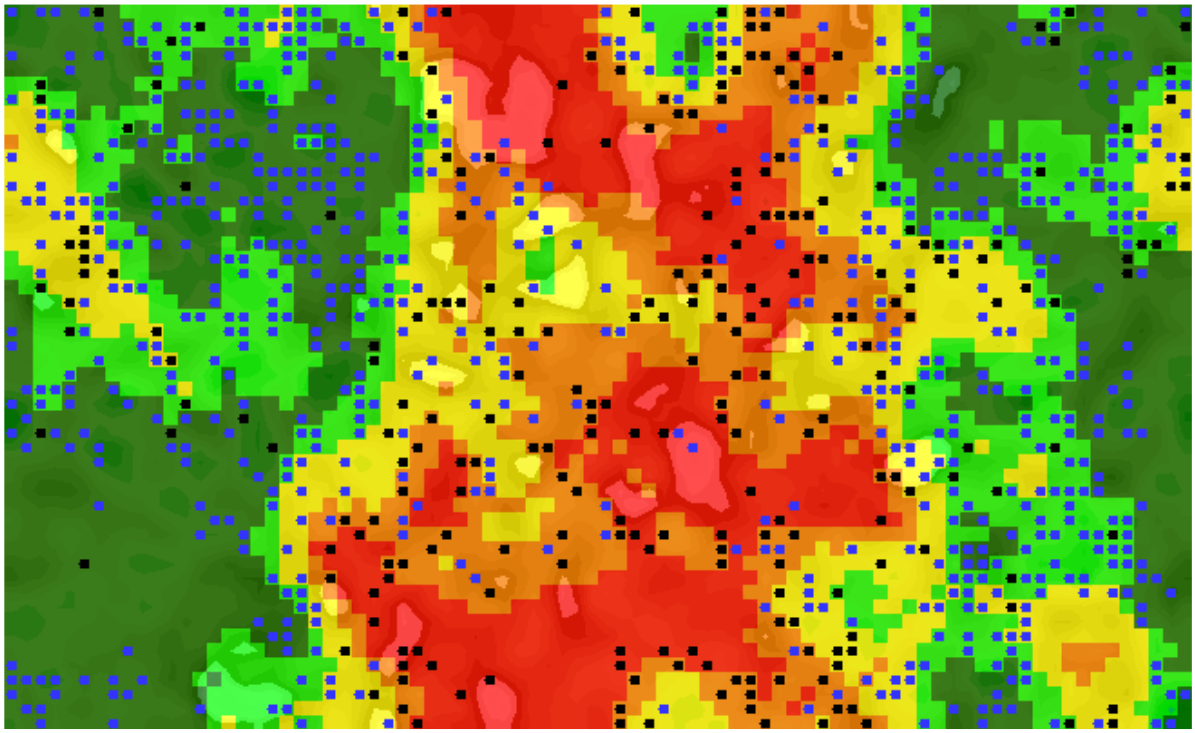

Fig. 8. ESOM risk analysis map for the year 2007 and remaining cases made visible

It was remarkable to observe that some of the remaining cases were located in the red area of the map, but were not classified by police officers as domestic violence. About 6.4% of the remaining cases were located in the red area of the map displayed in figure 8. About 22.1% of the cases located in the red area of the map were classified as non-domestic violence by police officers. In-depth analysis of these police reports revealed that the majority of these cases should have been classified as domestic violence. On the other hand, only a small percentage of the cases located in the dark green and green areas of the map were classified as domestic violence by police officers (4.8% and 12.4% respectively). Further scrutiny revealed that all of these cases actually described non-domestic violence incidents.

Table 9. Distribution of remaining cases of 2007 over different map areas

Domestic violence probability	Map area color	% of remaining cases located in map area	% classified as domestic violence	% classified as non-domestic violence
0-10%	dark green	28.1%	4.8%	95.2%
10-30%	green	30.0%	12.4%	88.6%
30-70%	yellow	21.5%	37.7%	62.3%
70-90%	orange	14.0%	64.3%	35.7%
90-100%	red	6.4%	77.9%	22.1%

Based on the map displayed in Figure 8, a correct label can be automatically assigned to 64.5% of the remaining cases of the year 2007 (i.e. the cases located in the dark green, green and red areas of the map). The other cases (i.e. the cases located in the yellow area of the map) have to be classified manually. A similar result was obtained for the cases of the year 2006 and 2008. Based on the comprehensible classification rules discovered during the knowledge discovery exercise, we developed a Tomcat-based system to assist analysts in their labeling of cases. The system is currently used as a stand-alone application by the data quality management team (i.e. the back office). The long term goal is to make it available to all police officers in the organization (i.e. the front office) to assist them in their labeling of cases.

The labeling process, as performed by the data quality management team, consists of a number of steps that are, to a large extent, automated by the newly introduced system. First, the user can select a set of police reports for labeling (e.g. all police reports from the month October 2008). Subsequently, the classification rules that were discovered during the exploration of the data are applied to the cases. When

a case comes in for labeling, the first step consists in verifying whether one of the domestic violence rules is satisfied. If this is the case, the case is classified as domestic violence. If not, it is verified whether one of the non-domestic violence rules is applicable. If this is the case, the case is classified as non-domestic violence. Otherwise, the case is left unclassified. The remaining cases are projected onto the ESOM risk analysis map based on the cases labeled with the FCA rules. Using the combination of the classification rules and the ESOM risk analysis map, 91.0 % of cases can be classified automatically and correctly. This is a major improvement compared to the past situation where each incoming case had to be dealt with manually.

7. Conclusions

In this paper, we proposed an approach to knowledge discovery from unstructured text using FCA and ESOM. The approach was framed within C-K theory (i.e. the design square) to provide a deeper understanding of the nature of the exploration process, a process that is essentially human-centered. With this paper we argued for the discovery capabilities of FCA and ESOM, acting as information browsers in the hands of human analysts. The tools were shown to help analysts proceed with knowledge expansion by progressively looping through the design square in an effective way. We demonstrated the method using a real-life case study with data from the Amsterdam-Amstelland police. The case focused on the problem of distilling concepts indicating domestic violence from the unstructured text in police reports. The data exploration for this case study resulted in several improvements to the way domestic violence cases are dealt with and reported on in practice. This included the implementation of an effective early case filter to identify cases that truly warrant in-depth manual inspection.

Intensive audits of the police databases revealed that many police reports tended to be wrongly classified as domestic or as non-domestic violence. Our approach was used to discover new features that better distinguish domestic from non-domestic violence cases resulting in higher classification accuracy and an improvement of the domestic violence definition. Additionally, we found some regularly occurring situations that were often wrongly labeled as non-domestic violence by police officers (e.g. lover boys). Even-

tually, we managed to build an accurate and comprehensible classifier that automatically assigns a correct label to more than 90% of incoming cases. Moreover, a large number of cases incorrectly classified in the past were detected and corrected thanks to this procedure.

Potentially, in future work one could investigate how iceberg lattices and alpha lattices could be used to prune the FCA lattices. One could also investigate the potential of conceptual scaling for improving scalability of the lattices which is however very labor-intensive.

Acknowledgments

The authors would like to thank the police of Amsterdam-Amstelland for granting them the liberty to conduct and publish this research. In particular, we are most grateful to Deputy Police Chief Reinder Doeleman and Police Chief Hans Schönfeld for their continued support. Jonas Poelmans is aspirant of the Research Foundation – Flanders.

References

[1] Hatchuel, A. (1996) Les theories de la conception, Paris.
[2] Hatchuel, A., Weil, B. (2002) La théorie C-K: fondements et usages d'une théorie unifiée de la conception. Proceedings of Colloque sciences de la conception, Lyon, 15-16 mars 2002.
[3] Hatchuel, A., Weil, B. (1999) Pour une théorie unifiéee de la conception. Axiomatiques et processus collectifs, 1-27 CGS Ecole des Mines/GIS cognition – CNRS, Paris 1999.
[4] Hatchuel, A., Weil, B. (2003) A new approach of innovative design: an introduction to C – K theory. Proceedings of ICED'03, august 2003, Stockholm, Sweden, pp. 14.
[5] Hatchuel, A., Weil, B., Le Masson, P (2004) Building innovation capabilities. The development of Design-Oriented Organizations: In Hage, J.T. (Ed), Innovation, Learning and Macro-institutional Change: Patterns of knowledge changes.
[6] Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M., 2004. Crime data mining: a general framework and some examples. IEEE Computer, April 2004.
[7] Ritter, H. (1999) Non-Euclidean Self-Organizing Maps, pp. 97–109. Elsevier, Amsterdam.
[8] Kohonen, T. (1982), "Self-Organized formation of topologically correct feature maps", Biological Cybernetics, Vol. 43, pp. 59-69.
[9] Ultsch, A., Moerchen, F. (2005) ESOM-Maps: Tools for clustering, visualization, and classification with Emergent SOM. Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany, No. 46.
[10] Ultsch, A., Hermann, L. (2005) Architecture of emergent self-organizing maps to reduce projection errors. In Proc. ESANN 2005, pp. 1-6.
[11] Ultsch, A. (2004) Density Estimation and Visualization for Data containing Clusters of unknown Structure. In proc. GfKI 2004 Dortmund, pp. 232-239.
[12] Ultsch, A. (2003) Maps for visualization of high-dimensional Data Spaces. In proc. WSOM'03, Kyushu, Japan, pp. 225-230.
[13] Ultsch, A., Siemon, H.P. (1990) Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. Proc. Intl. Neural Networks Conf., pp. 305-308.
[14] Van Hulle, M. (2000) Faithful Representations and Topographic Maps from distortion based to information based Self-Organization. Wiley: New York.

[15] Pednault, E.P.D. (2000) Representation is everything. Communications of the ACM, Vol. 43, no. 8
[16] Marchionini, G. (2006) Exploratory search: from finding to understanding. Communications of the ACM, Vol. 49, no. 4
[17] Poelmans, J., Elzinga, P., Viaene, S., Van Hulle, M., Dedene, G. (2009). Gaining insight in domestic violence with emergent self organizing maps. Expert systems with applications, 36(9), 11864-11874
[18] http://databionic-esom.sourceforge.net/
[19] Ultsch, A. (1999) Data mining and knowledge discovery with Emergent SOFMS for multivariate Time Series. In Kohonen Maps, pp. 33-46.
[20] Ananyan, S. (2002) Crime Pattern Analysis Through Text Mining. Proceedings of the Tenth Americas Conference on Information Systems, New York, August 2004.
[21] Keus, R., Kruijff, M.S. (2000) Huiselijk geweld, draaiboek voor de aanpak. Directie Preventie, Jeugd en Sanctiebeleid van de Nederlandse justitie.
[22] Ganter, B., Wille, R. (1999), Formal Concept Analysis: Mathematical Foundations. Springer, Heidelberg.
[23] Wille, R. (1982), Restructuring lattice theory: an approach based on hierarchies of concepts, I. Rival (ed.). Ordered sets. Reidel, Dordrecht-Boston, pp. 445-470.
[24] Priss, U. (2005), Formal Concept Analysis in Information Science, Cronin, Blaise (ed.), Annual Review of Information Science and Technology, ASIST, Vol. 40.
[25] Wille, R. (2002), Why can concept lattices support knowledge discovery in databases?, Journal of Experimental & Theoretical Artificial Intelligence, 14: 2, pp. 81-92.
[26] Stumme, G., Wille, R., Wille, U. (1998), Conceptual Knowledge Discovery in Databases Using Formal Concept Analysis Methods, In: J.M. Zytkow, M. Quafofou (eds.): Principles of Data Mining and Knowledge Discovery, Proc. 2 nd European Symposium on PKDD '98, LNAI 1510, Springer, Heidelberg, 1998, pp. 450-458.
[27] Stumme, G. (2002) Efficient Data Mining Based on Formal Concept Analysis. Lecture Notes in Computer Science Vol. 2453, Springer, Heidelberg, pp. 3-22.
[28] Stumme, G. (2002), Formal Concept Analysis on its Way from Mathematics to Computer Science. Proc. 10th Intl. Conf. on Conceptual Structures (ICCS 2002). LNCS, Springer, Heidelberg.
[29] T. van Dijk, Huiselijk geweld, aard, omvang en hulpverlening (Ministerie van Justitie, Dienst Preventie, Jeugdbescherming en Reclassering, oktober 1997).
[30] Brachman, R., Anand, T. (1996) The process of knowledge discovery in databases: a human-centered approach. In advances in knowledge discovery and data mining, ed. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy. AAAI/MIT Press.
[31] Brachman, R.J., Selfridge, P.G., Terveen, L.G., Altman, B., Borgida, A., Halper, F., Kirk, T., Lazar, A., Mc Guinnes, D.L. and Resnick, L.A. (1993) Integrated support for data archaeology. International Journal of In telligent and Cooperative Information Systems, 2: pp. 159-185.
[32] Smyth, P., Pregibon, D., Faloutsos, C. (2002) Data-driven evolution of data mining algorithms. Communica tions of the ACM, Vol. 45, no. 8.
[33] Fayyad, U., Uthurusamy, R. (2002) Evolving data mining into solutions for insights. Communications of the ACM, Vol. 45, no. 8.
[34] Cole, R.J. (2000) The management and visualization of document collections using Formal Concept Analysis. Ph. D. Thesis, Griffith University.
[35] http://www.politie-amsterdam-amstelland.nl/get.cfm?id=86
[36] Christopher, A. (1965) A city is not a tree. Architectural Forum, Vol 122, No 1, April 1965, pp. 58-62 (Part I) and Vol 122, No 2, May 1965, pp. 58-62 (Part II).
[37] Hollywood, J., Strom, K., Pope, M. (2009) Can Data Mining Turn Up Terrorists? OR/MS Today – February.
[38] Keim, D.A. (2002) Information visualization and visual data mining. IEEE transactions on visualisation and computer graphics. Vol. 8, No. 1.
[39] Eidenberger, H. (2004) Visual Data Mining. Proceedings SPIE Optics East Conference, Philadelphia 26-28 October. Vol. 5601, pp. 121-132.
[40] Thomas, J., Cook, K. (2005) Illuminating the path: research and development agenda for visual analytics. IEEE-Press.
[41] Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2009). A case of using formal concept analysis in combination with emergent self organizing maps for detecting domestic violence. In : Lecture Notes in Artificial Intelligence, Vol. 5633(XI), (Perner, P. (Eds.)). Industrial conference on data mining ICDM 2009. Leipzig (Germany), 20-22 July 2009 (pp. 402 p.).

[42] Eklund, P., Ducrou, J., Brawn, P. (2004) Concept lattices for information visualization: can novices read line diagrams? P. Eklund (ed.): ICFCA 2004, LNAI 2961, pp. 57-73. Springer-Verlag Berlin Heidelberg.
[43] Domingo,S., Eklund, P. (2005) Evaluation of concept lattices in a web-based mail browser. F. Dau, M.-L. Mugnier, G. Stumme (Eds.): ICCS 2005, LNAI 3596, pp. 281–294.Springer-Verlag Berlin Heidelberg.
[44] Correira, J.H., Stumme, G., Wille, R., Wille, U. (2003) Conceptual knowledge discovery – a human-centered approach. Applied artificial intelligence, 17: 281-302.
[45] Valtchev, P., Missaoui, R., Godin, R. (2004) Formal concept analysis for knowledge discovery and data mining: the new challenges. P. Eklund (ed.), ICFCA 2004, LNAI 2961, pp. 352-371. Springer-Verlag Berlin Heidelberg.
[46] Elzinga, P. (2006) Textmining by fingerprints. Onderzoeksrapport huiselijk geweld zaken. IGP project Activiteit 0504
[47] Stumme, G., Wille, R. (eds.) (2000) Begriffliche Wissenverarbeitung – methoden und anwendungen. Springer Heidelberg.
[48] Scheich, P., Skorsky, M., Vogt, F., Wachter, C., Wille, R. (1993) Conceptual data systems. In: O. Opitz, B. Lausen, R. Klar (eds.) Information and classification. Springer Berlin Heidelberg, pp. 72-84.
[49] Hereth, J., Stumme, G., Wille, U., Wille, R. (2000) Conceptual knowledge discovery and data analysis. In: B. Ganter, G. Mineau (eds.) Conceptual structures: logical, linguistic and computational structures. Proc. ICCS 2000. LNAI 1867, Springer, Heidelberg, pp. 421-437.
[50] Becker, K., Stumme, G., Wille, R., Wille, U. , Zickwolff, M. (2000) Conceptual information systems discussed through an IT-security tool. In: R. Dieng, O. Corby (eds.) Knowledge engineering and knowledge management. Methods, models and tools. Proc. EKAW '00. LNAI 1937, Springer, Heidelberg, pp. 352-365.
[51] Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2008). An exploration into the power of formal concept analysis for domestic violence analysis. In : Lecture Notes in Computer Science, 5077. Industrial Conference on Data Mining ICDM. Leipzig (Germany), 16-18 July 2008 (pp. 404-416). Springer.
[52] Poelmans, J., Elzinga, P., Viaene, S., Van Hulle, M., Dedene, G. (2009). How emergent self organizing maps can help counter domestic violence. World Congress on Computer Science and Information Engineering (CSIE). Los Angeles (USA), 31 March - 2 April 2009.
[53] Poelmans, J., Elzinga, P., Viaene, S., Dedene, G., Van Hulle, M. (2009). Analyzing domestic violence with topographic maps: a comparative study. 7th International Workshop on Self-Organizing Maps (WSOM). St. Augustine, Florida (USA), 8-10 June 2009.

Appendix A: Formal Concept Analysis

The starting point of the analysis is a database table consisting of rows M (i.e. objects), columns *F* (i.e. attributes) and crosses $T \subseteq M \times F$ (i.e. relationships between objects and attributes). The mathematical structure used to reference such a cross table is called a formal context (*M, F, T*). An example of a cross table is displayed in Table 1. Here, reports of domestic violence (i.e. the objects) are related (i.e. the crosses) to a number of terms (i.e. the attributes): a report is related to a term if the report contains this term. The dataset in Table 1 is an excerpt from the one we used in our research. Given a formal context, FCA then derives all concepts from this context and orders them according to a subconcept-superconcept relation, which results in a line diagram (a.k.a. lattice).

Table 1. Example of a formal context

	kicking	dad hits me	stabbing	cursing	scratching	maltreating
report 1	X	X				X
report 2			X	X	X	
report 3	X	X	X	X	X	
report 4						X
report 5				X	X	

The notion of concept is central to FCA. The way FCA looks at concepts is in line with the international standard ISO 704, which formulates the following definition. A concept is considered to be a unit of thought constituted of two parts: its extension and its intension [22, 23]. The extension consists of all objects belonging to the concept, while the intension comprises all attributes shared by those objects. Let us illustrate the notion of concept in a formal context using the data in Table 1. For a set of objects $O \subseteq M$, the common features, written $\sigma(O)$, can be identified with the following formula:

$$A = \sigma(O) = \{ f \in F \mid \forall o \in O : (o, f) \in T \}$$

Take, for example, the attributes that describe report 5 in Table 1. By collecting all reports of this context that share these attributes, we get a set $O \subseteq M$ consisting of reports 2, 3 and 5. This set *O* of objects is closely connected to set A, consisting of the attributes “cursing” and “scratching”.

$$O = \tau(A) = \{ i \in M \mid \forall f \in A : (i, f) \in T \}$$

In other words, O is the set of all objects sharing all attributes of A, and A is the set of all attributes that are valid descriptions for all the objects contained in O. Each such pair (O, A) is called a formal concept (or concept) of the given context. The set $A = \sigma(O)$ is called the intent, while $O = \tau(A)$ is called the extent of the concept (O, A).

There is a natural hierarchical ordering relation between the concepts of a given context that is called the subconcept-superconcept relation.

$$(O_1, A_1) \subseteq (O_2, A_2) \Leftrightarrow (O_1 \subseteq O_2 \Leftrightarrow A_2 \subseteq A_1)$$

A concept $d = (O_1, A_1)$ is called a subconcept of a concept $e = (O_2, A_2)$ (or equivalently, e is called a superconcept of a concept d) if and only if the extent of d is a subset of the extent of e (or equivalently, if and only if the intent of d is a superset of the intent of e). For example, the concept with intent "cursing", "scratching" and "stabbing" is a subconcept of a concept with intent "cursing" and "scratching". With reference to Table 1, the extent of the latter is composed of reports 2 and 3, while the extent of the former is composed of reports 2, 3 and 5.

The set of all concepts of a formal context combined with the subconcept-superconcept relations defined for these concepts gives rise to the mathematical structure of a complete lattice, called the concept lattice of the context. It is made accessible to human reasoning by using the representation of a (labeled) line diagram. The line diagram in Figure 2, for example, represents the concept lattice of the formal context abstracted from Table 1. The circles or nodes in this line diagram represent the formal concepts. The diagram displays only concepts that describe objects and is therefore a subpart of the concept lattice. The shaded boxes (upward) linked to a node represent the attributes used to name the concept. The non-shaded boxes (downward) linked to a node represent the objects used to name the concept. The information contained in the formal context of Table 1 can be distilled from the line diagram in Figure 2 by applying the following reading rule: an object "g" is described by an attribute "m" if and only if there is an ascending path from the node named "g" to the node named "m." For example, report 5 is described by the attributes "cursing" and "scratching."

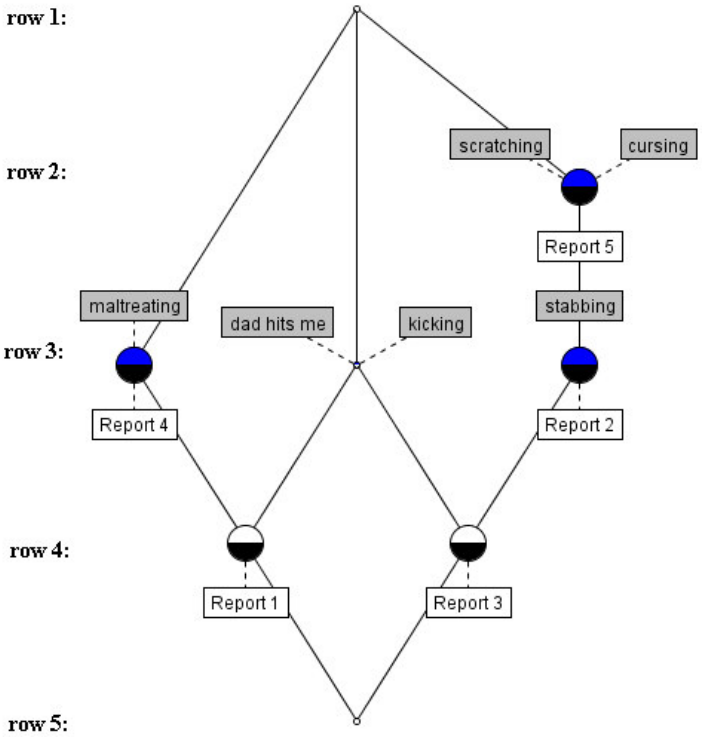

Fig. 2. Line diagram corresponding to the context from Table 1

Retrieving the extension of a formal concept from a line diagram such as the one in Figure 2 implies collecting all objects on all paths leading down from the corresponding node. In this example, the objects associated with the third concept in row 3 are reports 2 and 3. To retrieve the intension of a formal concept, one traces all paths leading up from the corresponding node in order to collect all attributes. In this example, the third concept in row 3 is defined by the attributes "stabbing," "cursing" and "scratching". The top and bottom concepts in the lattice are special: the top concept contains all objects in its extension, whereas the bottom concept contains all attributes in its intension. A concept is a subconcept of all concepts that can be reached by traveling upward and it will inherit all attributes associated with these superconcepts. Note that the extension of the concept with attributes "kicking" and "dad hits me" is empty. This does not mean that there is no report that contains these attributes. However, it does mean that there is no report containing only these two attributes.

Appendix B: Emergent SOM

An ESOM map is composed of a set of neurons *I*, arranged in a hex-grid map structure. A neuron $i \in I$ is a tuple (w_i, p_i) consisting of a weight vector $w_i \in W$ and a position $p_i \in P$ in the map. The input space $D \subset R^n$ is a metric subspace of R^n. The training set $E = \{x_1, ..., x_k\}$ with $x_1, ..., x_k \in R^n$ consists of input samples presented during the ESOM training. The training algorithm used is the online training algorithm in which the best match for an input vector is searched and the neighborhood of the map is updated immediately. When an input vector x_i is supplied to the training algorithm, the weight of a neuron $n_1 = (w_1, p_1)$ is modified as follows. Let $\eta \in [0,1]$, then

$$\Delta w_1 = r \times h \times (bm_i, n_1, r) \times (x_i - w_1)$$

the best-matching neuron of an input vector $x_i \in D$

$$D \to I : bm_i = bm(x_i)$$

is the neuron $n_b \in I$ having the smallest Euclidean distance to x_i;

$$n_b = bm(x_i) \Leftrightarrow d(x_i, w_b) \le d(x_i, w_b) \forall w_b \in W .$$

where $d(x_i, w_j)$ stands for the Euclidean distance of input vector x_i to weight vector w_j. The neighborhood of a neuron

$$N_i = N(n_i) = \{n_j \in M \mid h_{ij}(r) \neq 0\}$$

is the set of neurons surrounding neuron n_i and determined by the neighborhood function *h*. The neighborhood defines a lattice of neurons in the map space *K,* while *r* is called the neighborhood radius.

The map produced maintains the neighborhood relationships that are present in the input space and can be used to visually detect clusters. It also provides the analyst with an idea of the complexity of the dataset, the distribution of the dataset (e.g. spherical) and the amount of overlap between the different classes of objects. The lower-dimensional data representation is also an advantage when constructing classifiers. ESOM maps can be created and used for data analysis by means of the publicly available Da-

tabionics ESOM Tool [18]. With this tool the user can construct ESOMs with either flat or unbounded (i.e. toroidal) topologies.